Neural Conversation Generation with Auxiliary Emotional Supervised Models

Zhou, Guangyou

2019

Zhou , G , Fang , Y , Peng , Y & Lu , J 2019 , ' Neural Conversation Generation with Auxiliary Emotional Supervised Models ' , ACM Transactions on Asian Language Information Processing , vol. 19 , no. 2 , 19 . https://doi.org/10.1145/3344788

http://hdl.handle.net/10138/308003
https://doi.org/10.1145/3344788

other

Neural Conversation Generation with Auxiliary Emotional Supervised Models

GUANGYOU ZHOU*, School of Computer, Central China Normal University
YIZHEN FANG, School of Computer, Central China Normal University
YEHONG PENG, School of Computer, Central China Normal University
JIAHENG LU, Department of Computer Science, University of Helsinki

An important aspect of developing dialogue agents involves endowing a conversation system with emotion perception and interaction. Most existing emotion dialogue models lack the adaptability and extensibility of different scenes because of their limitation to require a specified emotion category or their reliance on a fixed emotional dictionary. To overcome these limitations, we propose a neural conversation generation with auxiliary emotional supervised model (nCG-ESM) comprising a sequence-to-sequence (Seq2Seq) generation model and an emotional classifier used as an auxiliary model. The emotional classifier was trained to predict the emotion distributions of the dialogues, which were then used as emotion supervised signals to guide the generation model to generate diverse emotional responses. The proposed nCG-ESM is flexible enough to generate responses with emotional diversity, including specified or unspecified emotions, which can be adapted and extended to different scenarios. We conducted extensive experiments on the popular dataset of Weibo post-response pairs. Experimental results showed that the proposed model was capable of producing more diverse, appropriate, and emotionally rich responses, yielding substantial gains in diversity scores and human evaluations.

Categories and Subject Descriptors: [**Artificial intelligence**]: Natural language processing–Natural language generation; *Emotional Conversation*

General Terms: Theory, Experimentation, Algorithms, Performance

Additional Key Words and Phrases: Neural Conversation, Sequence-to-sequence Model, Natural Language Processing

ACM Reference Format:
Zhou, G., Fang, Y., Peng, Y., Lu, J. 2018. Neural Conversation Generation with Auxiliary Emotional Supervised Models. ACM Trans. Asian Low-Resour. Lang. Inf. Process. V, N, Article A (January YYYY), 19 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Communication is a great driving force for the development and progress of human society, and emotions are an indispensable aspect of social communication. Emotional competencies are significant for social interactions, because emotions play an important role in conveying information about human's thoughts and intentions and coordi-

This work described in this article is supported by the National Natural Science Foundation of China (No. 61573163) and the Fundamental Research Funds for the Central Universities.
Author's addresses: G. Zhou, School of Computer, Central China Normal University, China; email: gyzhou@mail.ccnu.edu.cn; Y. Fang and Y. Peng, School of Computer, Central China Normal University, China; email: {yzfang, yhpeng}@mails.ccnu.edu.cn; J. Lu, Department of Computer Science, University of Helsinki, Findland; email: jiaheng.lu@helsinki.fi.
*Author for Correspondence: Guangyou Zhou, School of Computer, Central China Normal University, China.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

nating social encounters [Keltner and Haidt 2001]. Therefore, a good dialogue agent should have the ability to perceive and express emotions through interaction with humans.

Conversation systems are important to facilitate natural interactions between humans and virtual agents [Li et al. 2016a]; however, early systems were limited to a specific domain [Bohus and Rudnicky 2005], a pre-constructed database [Ji et al. 2014], or based on the framework of statistical machine translation (SMT) [Ritter et al. 2011]. Because the expanding technology of deep neural networks facilitates various applications in natural language processing (NLP), many researchers in both academia and industry are actively exploring the paradigm of neural conversation generation. Although considerable advances have been made regarding open-domain conversation generation using large-scale conversational data (e.g., Weibo, WeChat, or Twitter) [Ritter et al. 2011; Shang et al. 2015; Li et al. 2016b; Vijayakumar et al. 2016; Mou et al. 2016; Xing et al. 2017; Shao et al. 2017], the responses still tend to be dull, generic [Sordoni et al. 2015; Serban et al. 2016; Li et al. 2016a], and repetitive [Li et al. 2016c]. Recently, studies focused on improving content quality for conversation generation, including objective functions promoted with diversity information [Li et al. 2016a], beam search for diverse decoding [Vijayakumar et al. 2016; Shao et al. 2017; Li et al. 2016b], reinforcement learning [Li et al. 2016c], adversarial learning [Li et al. 2017], and topic information integration [Xing et al. 2017]. These efforts are meaningful; however, a chat bot is still unable to communicate with a user naturally when it lacks of the emotion perception and interaction.

Most recently, some promising studies that captured the importance of emotional factors have attempted to endow the responses with emotions; unfortunately, existing approaches, while pioneering, cannot successfully generate appropriate emotional responses based on the posts. Emotion Chatting Machine (ECM) proposed by Zhou et al. [2017] can only generate emotion specified replies, and Asghar et al. [2017] utilized three heuristic rules to build an affective neural dialogue generation model that relies upon an existing emotional dictionary, making it difficult to extend the method to other languages and datasets. In this article, we present a neural conversation generation with auxiliary emotional supervised models (nCG-ESM). The main idea is to make the dialogue generation system with emotional intelligence by providing emotional guidance during the learning process. In this system, the emotional distributions of the generated responses depend on the guidance signal from the emotional supervisor, which makes the system to learn the responses with different emotional distributions according to distinct emotional supervisory signals. We mainly focused on generating responses with five specified emotions and three unspecified emotions. Please note that our model can not only generate responses with these emotional distributions, but also can be easily extended to generate additional different emotional distribution responses to adapt to other scenes.

In a nutshell, nCG-ESM comprises two parts: the sequence-to-sequence (Seq2Seq) model [Zhou et al. 2017] to generate the reply and the auxiliary emotion classifier model for the responses over multiple emotional distributions. The emotion classifier is capable of producing an emotion distribution for each dialogue sentence, thereby assisting the neural conversation system in generating implied adaptive emotional responses (i.e., unspecified emotions) or designated specific emotional responses (i.e., specified emotions). Furthermore, the unspecified emotions contribute to generating adaptive emotional responses autonomously by utilizing original emotional information from the emotion distributions of source sentences, whereas specified emotions can provide different alternative emotional responses for various dialogue scenes. Therefore, nCG-ESM is adaptable and extendible in different context cases. Addition-

ally, to alleviate the problem of generating repeated words, we extended previous work [Lin et al. 2017] to augment a redundancy penalty term with loss functions.

Overall, our contributions in this article are shown as below:

— We presented a novel neural conversation generation with auxiliary emotional supervised model (nCG-ESM), which uses an emotional classifier as an auxiliary model to teach the Seq2Seq model to generate emotional responses.
— We developed several different variants of nCG-ESM model to generate different kinds of emotional responses, including five specified emotions and three unspecified emotions.
— We conducted extensive experiments on a Chinese emotional conversation dataset. Experimental results showed that the presented model was capable of generating diverse emotional responses. This model can be easily extended to diverse situations and additional datasets.

The rest of this article is organized as follows. Section 2 presents the related work and Section 3 describes the proposed nCG-ESM in detail. The exprimental evaluation and results are given in Section 4. Finally, we conclude the paper and present the possible future work in Section 5.

2. RELATED WORK

Conventional dialogue systems are based on rules or templates [Bohus and Rudnicky 2005; Williams and Young 2007] and are thereby reliant on substantial manual effort and limited to specific domains. With the increasing popularity of social media, large-scale data is available to solve conversation problems through data-driven methods, such as retrieval-based [Ji et al. 2014; Wang et al. 2013] and SMT-based methods [Ritter et al. 2011]. However, these two methods are either limited to an existing database or are only semantically equivalent to the original post. Recent success in the NLP field based on the use of neural networks, such as those involving language understanding [Mikolov et al. 2010], question answering [Zhou et al. 2015; Zhou and Huang 2017; Xie et al. 2017], neural machine translation [Bahdanau et al. 2014] and sentiment analysis [Zhou et al. 2016b], has inspired the researchers to employ neural network techniques to neural dialogue generation [Shang et al. 2015; Serban et al. 2015, 2016, 2017; Mou et al. 2016; Sordoni et al. 2015; Tian et al. 2017; Zhou et al. 2016a]. Shang et al. [2015] proposed a neural responding machine with an attention mechanism based on the general encoder-decoder framework, and Serban et al. [2015] extended a hierarchical neural network by utilizing historical information of conversations. Recently, Tao et al. [2017] proposed an evaluation methodology for open-domain dialogue systems considering both the ground truth and its query.

Additionally, considerable work in conversation generation has improved the quality and increased the diversity of conversation content, but through different methods, including content-introducing approaches [Mou et al. 2016; Xing et al. 2017], diversity-promoting objective functions promoted with diverstiy information [Li et al. 2016a], beam search with diverse decoding [Vijayakumar et al. 2016; Shao et al. 2017; Li et al. 2016b], reinforcement learning [Li et al. 2016c], adversarial learning [Li et al. 2017], and hybrid approaches [Yan et al. 2017; Ghazvininejad et al. 2017]. Li et al. [2016a] defined an objective function using maximum mutual information rather than traditional maximum-likelihood estimation, which tends to generate safe responses, and Ghazvininejad et al. [2017] integrated retrieved context-relevant facts with the conversation history to build a knowledge-enhanced neural conversation system.

Apart from considering the syntax and semantics of responses, a good dialogue agent should also be able to perceive and express emotions. In order to build a neural conversation system at the human level, some methods have been proposed to endow dia-

logue systems with emotions, although these have relied upon rule-based conversation methods to incorporate emotion factors which is not scalable to large datasets [André et al. 2004; Skowron 2009; Skowron et al. 2011; Ptaszynski et al. 2009]. Recent studies attempted to address the emotion response generation problem by using large-scale data-driven methods and Seq2Seq frameworks [Zhou et al. 2017; Asghar et al. 2017; Zhang et al. 2017]. Asghar et al. [2017] proposed a novel method to generate emotional responses based on affective word embeddings, affective loss functions, as well as the diverse beam search. Sun et al. [2017] addressed the emotional factors of response generation by using a series of input transformations. ECM [Zhou et al. 2017] is a Seq2Seq model that integrates emotional factors, including an emotion embedding as well as an internal memory and an external memory with generated responses, while Zhang et al. [2017] employed a Seq2Seq learning framework under a multi-task manner to build an emotional dialogue system. Additionally, previous studies attempted to employ hybrid approaches rather than pure neural network-based methods to build a neural emotional dialogue system, such as the ensemble framework by integrating retrieval- and generation-based conversation systems [Zhuang et al. 2017].

However, to our knowledge, autonomous and adaptive emotional response generation has not been addressed. Previous studies primarily aimed to generate emotional responses based on several specified emotion categories [Zhou et al. 2017; Zhuang et al. 2017; Zhang et al. 2017; Sun et al. 2017] or reliance on emotional dictionaries [Asghar et al. 2017]. Recently, Peng et al. [2019] proposed a topic-enhanced emotional conversation generation system by using an attention mechanism with the dynamic setting to automatically acquire the task-specific information. The proposed system obtained the topic information using a Twitter LDA instead of the auxiliary emotional supervised model. There are three salient features in our work: 1) our model uses emotion distributions from the emotion classifier, which eliminates the constraints of an emotional dictionary; 2) our model is not restricted to specified emotions, but can generate adaptive emotional responses according to the implications of posts; and 3) our proposed approach can be easily adapted to generate more emotional distribution responses.

3. NEURAL CONVERSATION MODEL WITH AUXILIARY EMOTIONAL MODELS

In this section, we describe a general framework of the proposed emotional conversation system and present the each component of our model in details.

Our goal is to endow conversation agents with emotional intelligence by using a flexible and extensible approach. Therefore, given a post $X = (x_1, x_2, \ldots, x_{T_x})$, the model can generate an emotional response $Y = (y_1, y_2, \ldots, y_{T_y})$. To this end, we developed a neural conversation generation with auxiliary emotional supervised model (nCG-ESM). Our system incorporates emotion distribution into an encoder-decoder framework that utilizes an emotion classifier to supervise the learning process of the neural conversation model. As shown in Fig. 1 and Fig. 2, we considered two situations of emotional response generation: specified emotion and unspecified emotion. Specifically, we developed several variants of models to generate responses with different emotional distributions that could be classified into two groups:

(1) In the unspecified emotion situation, we used three heuristics to generate diffrent emotional supervisory signals, enabling the model to generate different emotional responses based on emotion distributions of the dialogue without human intervention (Fig. 1).
(2) In the specified emotion situation, the model is infused with specified emotions, including {*Angry, Disgust, Happy, Like, Sad*} [Zhou et al. 2017] , to generate emotional responses (Fig. 2).

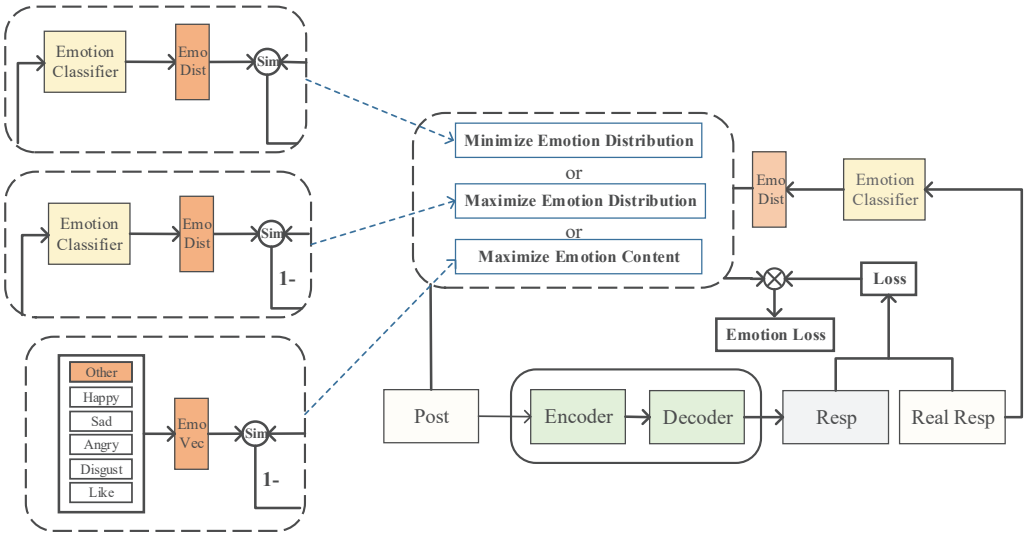

Fig. 1: Overview of the unspecified emotion model in nCG-ESM. The unspecified emotion model consists of three variants, namely, Minimize Emotion Distribution (MinDis), Maximize Emotion Distribution (MaxDis), and Maximize Emotion Content (MaxEmo). The right part of Fig.1 shows the master of model, of which the dashed line part can be replaced by the three variants in the left part of Fig.1 separately.

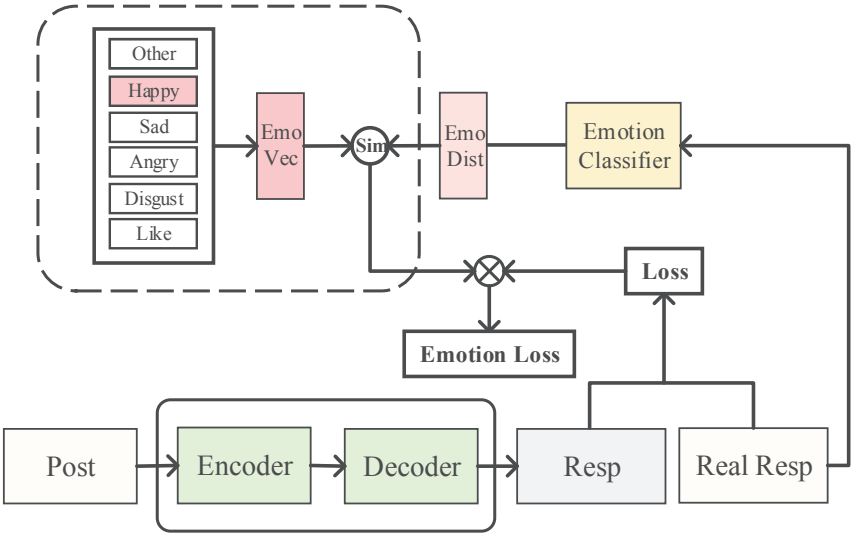

Fig. 2: Overview of the specified emotion model in nCG-ESM. There are five emotional responses that can be specified, with the example provided specifying happy emotion.

3.1. The Preliminary Seq2Seq Model

Our system adopted an optimized encoder-decoder framework, as previously described in [Sutskever et al. 2014]. The encoder reads the input sequence X and encodes the input as a contextual vector c via a RNN, and then the decoder predicts the conditional probability of Y with the contextual vector as the input [Xing et al. 2017]. We adopted a BiLSTM for the encoder RNN, including a forward and a backward LSTMs. The forward LSTM encodes the input sequence X as series of hidden states $\vec{h} = (\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_{T_x})$ in a forward direction, while the backward LSTM converts the reverse input sequence to another hidden states $\overleftarrow{h} = (\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_{T_x})$. Finally, we concatenate the two directional states to form the final representation $h_j = [\overrightarrow{h}_j; \overleftarrow{h}_j]$ at time step j. The formulation is written as:

$$h_j = \mathrm{BiLSTM}(h_{j-1}, x_j) \tag{1}$$

During the decoding phase, we use a unidirectional LSTM. Thus, the state s_i at time i is calculated as

$$s_i = \mathrm{LSTM}(s_{i-1}, y_{i-1}, c_i) \tag{2}$$

where y_{i-1} is the word embedding of a previously decoded result, and c_i is the context vector which is distinct for each token y_i, and computed as follows:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{3}$$

$$\alpha_{ij} = \frac{\exp(\eta(s_{i-1}, h_j))}{\sum_{k=1}^{T_y} \exp(\eta(s_{i-1}, h_k))} \tag{4}$$

where η is a multilayer perceptron. The loss function used by Seq2Seq model is typically represented by the maximum-likelihood estimation (MLE) objective function defined as follows:

$$L_{mle}(\theta) = -\log p(Y|X) = -\sum_{i=1}^{T_y} \log p(y_i|y_1, y_2, \ldots, y_{i-1}, X) \tag{5}$$

The next token y_i is generated by sampling from the output probability distribution d_i, denoted as follows:

$$\begin{aligned} y_i \sim d_i &= P(y_i|y_1, y_2, \ldots, y_{i-1}, X) \\ &= \mathrm{softmax}(W_d s_i + b) \end{aligned} \tag{6}$$

where W_d and b are parameters in the output layer.

3.2. Auxiliary Emotional Supervised Models

Here, we assumed that each dialogue sentence had its own emotion distribution containing the abstract emotional information. This assumption stemmed from the fact that each daily conversational text always has one or several emotional tendencies. Based on this assumption, we used an emotion classifier to produce emotion distributions that could guide our model to generate emotional responses. In Section 4.1.2, our experimental results showed that the emotion classifier of BiLSTM had the best performance. Therefore, we choose a BiLSTM emotion classifier as the auxiliary model for

our conversation system. We used $S = (s_1, s_2, \ldots, s_T)$ as an input sequence to formulate the process as follows:

$$h_i = \mathrm{BiLSTM}(h_{i-1}, s_i) \tag{7}$$

$$V_i = U\tanh(\mathrm{W_V} h_i) \tag{8}$$

$$\beta_i = \frac{\exp(V_i)}{\sum_{j=1}^{n} \exp(V_j)} \tag{9}$$

$$c_i = \sum_{i=1}^{T} \beta_i h_i \tag{10}$$

where h_i is the hidden vector, c_i is the context vector, U and W_V are parameters, and β is the weight of attention.

Instead of using the maximum probability label as the ultimate target classification for each sentence, we retained the output from the final fully connected softmax layer as the emotion distribution, which could be rewritten as a probability distribution over emotion labels:

$$EmoDist_S = \mathrm{softmax}(\mathrm{W}\boldsymbol{c} + b) \tag{11}$$

Note that the purpose of our emotion classifier differs from that described previously [Zhou et al. 2017]. The latter is used to annotate the single emotion category for each dialogue sentence and specify the fixed emotions for responses. However, in practice, one dialogue sentence possibly contains have multiple emotions. Therefore, in this work, we propose an emotion distribution to capture all emotional information for emotional dialogue generation.

3.3. Emotional Objective Functions

Since we can obtain the emotion distributions of every post and response through the emotion classifier, we further designed different emotional Seq2Seq generation models, which enable the automatic generation of emotional responses in both specified and unspecified manners.

3.3.1. Unspecified Emotion. In general, the emotional tendency of a short conversation between the post and the response tends to be consistent. To infuse the response with the emotional information consistent with that of the given post, we presented a Minimize Emotion Distribution variant (Fig. 1). Specifically, we first feed the post and the ground truth response to the emotional classifier to obtain corresponding emotion distributions, $EmoDist_X$ and $EmoDist_Y$, and then compute their similarity according to cosine distance. We then multiply the MLE objective function by the cosine value as follows:

$$L_{MinDis}(\theta) = -cos(EmoDist_X, EmoDist_Y) \log p(Y|X) \tag{12}$$

In addition to remaining aligned with emotions associated with dialogue applicable to most situations, the model is also occasionally required to transform the emotional direction of the dialogue. In situations where the conversation is deadlocked, changing

the emotional direction of the conversation is an effective way to avoid embarrassment. From the perspective of training an open-domain dialogue system, inconsistent emotional responses are more interesting and more useful when regulating the overall atmosphere of a conversation. As described in Fig. 1, we designed a Maximize Emotion Distribution variant to maintain inconsistencies between the emotions in the generated response and the given post. Compared to Eq. (12), we normalized the MLE objective function by multiplying one minus the cosine value of the maximization of emotion dissonance:

$$L_{MaxDis}(\theta) = -(1 - cos(EmoDist_X, EmoDist_Y)) \log p(Y|X) \tag{13}$$

Additionally, we considered another situation where the emotional connections with posts are ignored, and the system can spontaneously generate responses involving plentiful emotions. Comprehensibly, the utterances of outgoing and passionate people in daily life are often more affective; therefore, we designed a Maximize Emotion Content variant (Fig. 1) and redesigned the objective function as follows:

$$L_{MaxEmo}(\theta) = -(1 - cos(EmoVec_{other}, EmoDist_Y)) \log p(Y|X) \tag{14}$$

where the $EmoVec_{other}$ is the $Other$ emotion vector. Our goal was to separate $EmoDist_Y$ (i.e., the emotion distribution of the response) as far away from the $EmoVec_{other}$ as possible in the space.

3.3.2. Specified Emotion. Note that the specific emotional responses are also required in some scenarios, as in cases where certain responses are required to maintain positive emotion in the service industry. Furthermore, it is difficult to determine the most appropriate emotional response, given that the emotion of the responses from different people is highly subjective, even in the same conversation. Various candidate emotions provide multiple emotional responses, making the system easy to extend upon selection of the most appropriate response.

From Fig. 2, our proposed system can specify five different emotion categories, including {*Angry, Disgust, Happy, Like, Sad*} [Zhou et al. 2017], to generate different emotional responses. We represented the various emotion categories as different one-hot emotion vectors (i.e.,$EmoVec$), with the emotional classifier capable of yielding emotion distributions of the ground-truth responses. We then evaluated the emotional dialogue interaction under the specified emotion state by minimizing the distance between the specified emotion vector and the emotion distribution of the response, thereby guiding the system to generate the expected response consistent with the specified emotion. As mentioned, we used cosine function to measure the disparity between the emotion distribution and the specified emotion vector, thereby obtaining the objective function of the specified emotion model, denoted as follows:

$$L_{SpeEmo}(\theta) = -cos(EmoVec, EmoDist_Y) \log p(Y|X) \tag{15}$$

where $EmoVec$ is one of the five specified emotion vectors, and $EmoDist_Y$ is the emotion distribution of the ground-truth response.

3.4. Penalization Term

Many previous studies based on the Seq2Seq framework made considerable advances in open-domain conversation generation using the MLE objective, whereas the generated responses tend to be repetitive (e.g., no,no,no,no,no). The repetivie words cannot meet the requirements of a good dialogue system. To address this problem, we used a penalization term to encourage the disparity of the probability distributions

of the output words across different time steps in the annotation softmax output. Formally, the probability distribution matrix of each output response is denoted as $D = (d_1, d_2, \ldots, d_n) \in \mathbb{R}^{n*v}$, where d_i is the probability distribution at time step i, n and v represent the length of the response and the size of the vocabulary, respectively.

The Kullback Leibler (KL) divergence is one option for assessing diversity between any two probability distributions, such as d_i and d_j; however, we want each probability distribution to focus on the true word of the ground truth, which is not a feature of KL divergence. Therefore, we extended the redundancy penalty term, which overcame the shortcoming of KL divergence to measure the disparity between different probability distributions. Following the literature [Lin et al. 2017], we define the penalization term as follows:

$$P = \left\| \left(DD^T - I \right) \right\|_F^{\ 2} \tag{16}$$

where I is an identity matrix, D^T is the transpose of matrix D, and $\|\cdot\|_F$ represents the F-norm of a matrix. We added this penalization term P to all of the emotional loss functions and minimized it together with them. Therefore, the model encourages every probability distribution to focus on the true word when forces the elements on the diagonal of DD^T to approximate 1, and increases the disparity between probability distributions when forces other elements to be 0, so that the generated words can be more different rather than repetitive.

4. EXPERIMENTS

In this part, we describe the experimental results to validate the effectiveness the proposed method. First, we introduce the datasets and parameter settings of our models, and then describe the compared methods and evaluation methodology. Furthermore, we show and analyze the experiment results including the automatic evaluation and human evaluation results in detail, and present some sample responses of different models for comparison.

4.1. Datasets

4.1.1. Dialogue Data. Here, we used the dialogue data from the Emotional Conversation Generation task of NLPCC 2017 Shared Tasks[1]. This dialogue data consists of more than 1 million post-response pairs, in which each post-response pair is labeled with an emotion label as well as an emotion category [Peng et al. 2019], and we removed the emotional labels to obtain the raw text of post-response pairs.

In order to achieve a good results, we designed a fine-grained vocabulary for our system to reduce the OOV problem. Specifically, we first constructed the initial vocabulary by choosing the words with a frequency > 80, and then adding the characters into vocabulary by splitting the OOV words. Therefore, the final vocabulary contained 12,819 entities, including high-frequency words and characters appearing > 80 times along with an END token and an UNK label. Additionally, we filtered extremely short post-response pairs and limited the sentences with less than 20 words. Finally, we retained about 85% of post-response pairs after data cleaning.

4.1.2. Emotion Classifier Data. We trained the emotion classifier on the data from the Emotion Analysis in Chinese Weibo Text task of NLPCC 2014 Shared Tasks [2]. The emotion classification in Chinese Weibo text includes eight manually annotated emotion categories {*Angry, Disgust, Fear, Happy, Like, Sad, Surprise, Other*}. We first removed the two infrequent categories {*Fear, Surprise*} and deleted the @ symbol and

[1] http://tcci.ccf.org.cn/conference/2017
[2] http://tcci.ccf.org.cn/conference/2014

the user name after it[3], followed by filtering of data using the method described for the dialogue data. We obtained 22,461 data entries for six categories {*Angry, Disgust, Happy, Like, Sad, Other*} [Zhou et al. 2017]. Please note that the proposed approach may depend on the accuracy of the emotional classifiers. If the classifiers fail to predict the emotional labels, the proposed method cannot generate the satisfactory emotional responses. In this paper, we trained several emotion classifers, including CNN, LSTM, and BiLSTM, with filtered dataset. The accuracy of various emotion classifiers are shown in Table I. From Table I, BiLSTM model achieved the best classification result among these classifiers, the similar results had also been elaborated in [Zhou et al. 2017].

Table I: The accuracy of different emotion classifiers.

Models	Accuracy
CNN	0.579
LSTM	0.593
BiLSTM	0.615

4.2. Parameter Settings

We used the parameter settings outlined in a previous study [Sutskever et al. 2014]. Specifically, our model used a BiLSTM encoder and an LSTM decoder, both with four layers, with each layer comprising 1,000 hidden units. We employed the 1,000 dimensional word embeddings using word2vec [Mikolov et al. 2013]. The vocabulary size was set to 12,819. Additional parameter settings are given below.

— We initialized the parameters with a uniform distribution between [-0.1, 0.1].
— We optimized the proposed method using Adam algorithm [Kingma and Ba 2014] with an initial learning rate of 0.001.
— The batch size was set to 128.
— We used gradient clipping (threshold: 5) to avoid gradient explosion.
— According to Li et al. [2017], we used a weighted loss that considered the different tf-idf scores for tokens within the responses to ensure less generic responses.
— We set the dropout rate to 0.2.

We implemented the model on the Tensorflow platform[4] according to the following training settings for each model.

— We pre-trained a attention-based Seq2Seq model [Sutskever et al. 2014; Hochreiter and Schmidhuber 1997] with an attention mechanism using cross-entropy loss for 20 epochs to decrease syntactical errors and improve convergence speed.
— Based on the pre-trained model, we continued to train the proposed system with every different loss function described in section 3.3 for five epochs, respectively.

[3]Adding user name after @ means to remind someone to follow this Weibo text, we denote as @XXX. Since @XXX exists a lot in this dataset, we filter out them to improve the data quality.
[4]`https://github.com/tensorflow/tensorflow`

4.3. Compared Methods

We implemented and compared with two methods: (1) We implemented ECM [Zhou et al. 2017], which has been integrated three mechanisms containing an emotion embedding, an internal memory, and an external memory for generating specified emotional replies. We used it to compare with the specified emotion model of nCG-ESM. Since there is no open source code for ECM, we developed the implementation by ourselves. (2) We compared with the Seq2Seq [5] model and the corresponding ground truth set. Previous conversation systems mainly used Seq2Seq as a basic comparator. We considered that the dialogue task differed from other NLP tasks using fixed and commonly accepted evaluation metrics; therefore, the ground truth has sufficient potential reference values for exploration. Intuitively, Seq2Seq methods are deemed as the minimal standards of conversation generation, whereas the ground truth represents an upper bound. Compared with Seq2Seq, we can observe the improvements in our models, and further, we can observe the extent of the distance between our models and the ground truth. To our knowledge, it is the first work to adopt the ground truth as a comparator for emotional dialogue system evaluation.

4.4. Evaluation Metrics

A previous study used automatic and manual evaluation methods [Zhou et al. 2017] simultaneously, whereas others only employed human judgement for evaluation [Asghar et al. 2017]. There is no uniform and accepted automatic evaluation technique for dialogue systems. Some adopted automatic metrics, such as BLEU and perplexity, were borrowed from adjacent research fields and are not entirely suitable for the evaluation of dialogue systems. Therefore, we chose the diversity of generated responses as our automatic metric, which we considered appropriate based on previous work [Li et al. 2016a], and designed a variety of methods for manual evaluation according to the specific cases.

For the automatic metric, we used *distinct-1* and *distinct-2*, as previously described in [Li et al. 2016a; Peng et al. 2019]. Additionally, we also explored two settings for human evaluation: the first adopted metrics used in [Zhou et al. 2017], and the second described in [Li et al. 2016a]. For the first setting, three annotators with rich Weibo experience were asked to score emotion responses based on content and emotion. Content (rating scale: 0, 1, 2) measures if the response is acceptable and fluent as a natural language sentence, and estimates if the logic structure of the response is appropriate [Shang et al. 2015]. In terms of emotion, there was a difference between the specified and unspecified emotion models because of their different emotional loss functions; therefore, we designed different evaluation criteria for emotion. For the specified emotion model, emotion (rating scale: 0 or 1) was used to judge if the emotion category of the generated responses were the same as the specified emotion one, and for the unspecified emotion model, emotion was used to evaluate whether the emotion of a response was suitable for the post.

The second setting designed exclusively for the unspecified emotion model was the preferred test, with three annotators employed to label their preference among the three variants of the unspecified emotion model and the two comparators in pairs. Ties were allowed. In this setting, we observed that different annotators may lead to different results.

[5]The parameter setting of Seq2Seq is the same as those of our models.

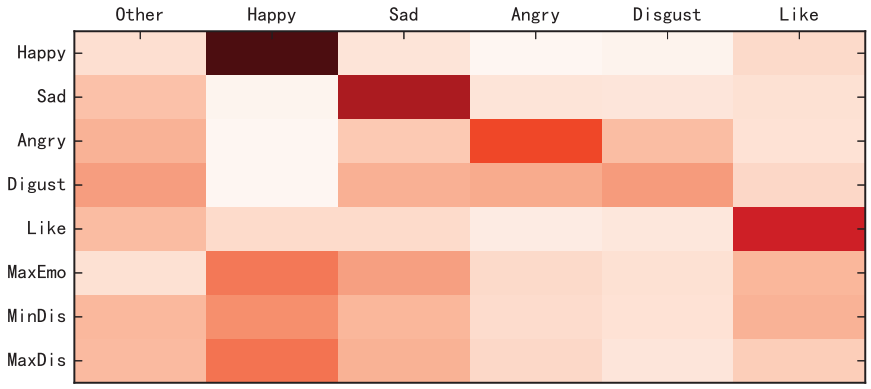

Fig. 3: Visualization of the emotion distributions of our models. The heat map indicates distributions over different emotions (top) of the corresponding models (left). Darker colors represent higher probabilities.

4.5. Experimental Results

4.5.1. Automatic Evaluation. We evaluated the generated responses from our models at the content level. The comparators contained the baselines (Seq2Seq and ECM) and the ground truth (Gold), as mentioned in Section 4.3.

Table II shows the automatic evaluation results. From the diversity metrics shown in Table II, we noted that our models achieved a substantial improvement in performance. Compared with Seq2Seq, the performance of specified and unspecified emotion models improved significantly for both *distinct-1* and *distinct-2*, and the unspecified emotion model exhibited a $>$ 100% jump in *distinct-1* and a 160% increase in *distinct-2*. In contrast with the ground truth, our models showed minimal differences in unigram diversity, indicating that our neural system learned well based on the diversity of individual tokens for response generation. And, as expected, both *distinc-1* and *distinct-2* of our model are higher than that of ECM, indicating that our models are more excellent in terms of diversity of words in the generated sentences. Table II shows that after adding penalization term to all the emotion loss functions, our system was more likely to generate different and informative words rather than repetitive and dull expressions, resulting in a greater diversity in the generated responses.

Table II: Performance of specified and unspecified emotion variants versus comparators upon automatic evaluation regarding different metrics.

Models	Specified / ECM						Unspecified				Comparator	
	Angry	Disgust	Happy	Like	Sad	Avg	MaxEmo	MaxDis	MinDis	Avg	Gold	Seq2Seq
distinct-1	0.102 / 0.042	0.107 / 0.044	0.107 / 0.056	0.110 / 0.064	0.103 / 0.042	**0.106** / 0.050	0.117	0.124	0.117	**0.119**	0.131	0.055
distinct-2	0.431 / 0.146	0.456 / 0.181	0.435 / 0.172	0.450 / 0.225	0.429 / 0.142	**0.440** / 0.173	0.467	0.489	0.481	**0.479**	0.700	0.184

Additionally, we analyzed the emotion distribution of each model. Specifically, for each model, we predicted the emotion categories of the generated responses using an emotion classifier and computed the probability of each emotion category. As shown in

Fig. 3, our specified emotion models had ability to generate the specified emotion responses, while $Angry$ and $Disgust$ models were weak to generate the specified emotion responses due to the insufficient training data, which makes the errors caused by the emotion classifier [Zhou et al. 2017]. Moreover, the color of $Other$ column in MaxEmo model was lighter than others, indicating that MaxEmo model was capable of generating richer and more colourful emotional responses. Since the emotion distributions of the MinDis and MaxDis models were related to the posts, the distributions of the emotion categories shown in the Fig. 3 are more average.

4.5.2. Human Evaluation. As noted, we explored two settings for human evaluation. For the first setting, 100 posts were randomly sampled from test data, and we generated responses for each post according to baselines (Seq2Seq and ECM) and the proposed model (nCG-ESM), including eight variants. Then we let the annotators to judge the content ($Cont.$) and emotion ($Emo.$) of all the responses according to human evaluation metrics. The annotation results are shown in Table III and Table IV.

Table III shows the content and emotion scores of the specified emotion models and ECM. In Table III, the average content score of our specified emotion models is higher than that of ECM. More specifically, the content scores of our models outperformed those of ECM in every emotion category, indicating that our models generated more natural and diverse responses after incorporating penalization term and emotion factors. And the average emotion score of ours performed a bit better than ECM. Although the emotion scores associated with $Angry$ and $Disgust$ were slightly lower than the average because of the lack of training data for these two categories, experimental results showed that they were recognized by the annotators at the emotion level. Additionally, we evaluate the agreements of labeled results using Fleiss' kappa [Fleiss 1971]. The agreement values for content and emotion at 0.48 and 0.63, respectively.

Table III: Human evaluation results of specified emotion variants and ECM.

Models	Specified / ECM					
	Angry	Disgust	Happy	Like	Sad	Avg
$Cont.$	1.36 / 1.15	1.49 / 1.2	1.49 / 1.3	1.56 / 1.29	1.4 / 1.16	**1.46** / 1.22
$Emo.$	0.46 / 0.43	0.44 / 0.41	0.69 / 0.85	0.55 / 0.49	0.54 / 0.47	**0.54** / 0.53

The human evaluation results of the unspecified emotion variants are shown in Table IV. Our models significantly outperformed the baseline for both metrics ($p < 0.005$ for $Cont.$ and $p < 0.05$ for $Emo.$ using t-test). After incorporation of emotion distributions and a penalization term, the performance of our unspecified emotion variants in $Cont.$ improved relative to that of baseline, implying that our models improved the content quality of the generated responses. Moreover, these three unspecified emotion

Table IV: Content and emotion scores based on human evaluation of unspecified emotion variants.

Models	MaxEmo	MaxDis	MinDis	Overall	Seq2Seq
$Cont.$	1.42	1.3	1.41	1.38	1.17
$Emo.$	0.64	0.63	0.67	0.65	0.47

Table V: The percent improvement by the unspecified emotion variants and Seq2Seq models over the ground truth based on pairwise human judgements.

Preference	MaxEmo	MaxDis	MinDis	Seq2Seq
Gold	29%	27%	**35**%	18%

Table VI: The percent improvement by the unspecified emotion variants over the Seq2Seq model based on pairwise human judgements.

Models	Seq2Seq-lose	Seq2Seq-win	Tie
MaxEmo	44%	28%	28%
MaxDis	52%	24%	24%
MinDis	56%	24%	20%

variants outperformed Seq2Seq in terms of emotion, indicating that the emotion loss functions contributed to generating emotional responses.

For the second setting, we used the same human evaluation data as that for the first setting to obtain preferences between the ground truth (Gold) and our unspecified emotion variants. Table V shows the percent gains by our models and baseline over the ground truth based on pairwise human judgments. All of the gains by our models were higher as compared with Seq2Seq, indicating that our system was more likely to generate emotional responses indistinguishable from humans.

Annotators were asked to state a preference between the unspecified emotion variants and Seq2Seq [Li et al. 2016c] by selecting their preferred responses. Their instructions included preferring responses that were more relevant to the post and more affective, rather than those that were more generic and repetitive. As shown in Table VI, compared with Seq2Seq, all three variants of the unspecified emotion model showed improved performance about 50% and worse performance $< 20\%$. These results suggested that the proposed unspecified emotion models could be better able to generate emotional responses at the human level.

4.5.3. Results Comparison. Fig. 4 shows sample responses generated by different models and used to compare our models with baselines (Seq2Seq and ECM) and the ground truth to enhance intuitive understanding. As illustrated in Fig. 4, for the second post is “I’m sick of this breakfast”, Seq2Seq model just generates the general response “Me too, It’s my first time to eat, haha” while our specified emotion model can generate responses with specified emotions, for example, sad response contains the expression of “not feeling well”. Moreover, our unspecified emotion model can generate more informative and emotional expressions related to the post, such as “too sweet”, “eaten too much”, and “hot and sour rice noodles”. Our results showed that the responses generated by our models were more affective and informative as compared with those generated by Seq2Seq, and some are also better relative to ground-truth responses. These results again validated that our model was capable of generating appropriate emotional responses that were either specified or unspecified.

5. CONCLUSIONS AND FUTURE WORK

In this study, we presented an nCG-ESM system to generate responses with a diversity of specified and unspecified emotions that is flexible enough to allow extension to situations involving a broader range of emotions. We enhanced the general conversation generation framework (i.e., the encoder-decoder system) by introducing a novel emotional supervised mechanism. Additionally, an emotional classifier was utilized to generate emotion distributions of the dialogue, which were then used to obtain emotional supervisory signals. Moreover, we improved the objective function by incorporating a redundancy penalty to avoid the generation of repeated words. The experimental results showed that our model not only generated better responses but also infused those responses with different emotion distributions.

Exciting followup research can be centered around the exploration of the neural emotional conversation system using reinforcement learning, where the main challenge is to make rational and acceptable decisions about the emotions of subsequent responses in various specific contexts. Another area of interest is to investigate multi-turn emotional conversation models which need to handle the problem of tracking and analyzing the emotions of historical dialogue. Additionally, we will conduct more experiments for different language datasets.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 61573163) and the Fundamental Research Funds for the Central Universities.

REFERENCES

Elisabeth André, Matthias Rehm, Wolfgang Minker, and Dirk Bühler. 2004. Endowing Spoken Language Dialogue Systems with Emotional Intelligence. In *Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings*. 178–187. https://doi.org/10.1007/978-3-540-24842-2_17

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2017. Affective Neural Response Generation. *CoRR* abs/1709.03968 (2017). arXiv:1709.03968

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). arXiv:1409.0473

Dan Bohus and Alexander I. Rudnicky. 2005. A principled approach for rejection threshold optimization in spoken dialog systems. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. 2781–2784.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A Knowledge-Grounded Neural Conversation Model. *CoRR* abs/1702.01932 (2017). arXiv:1702.01932

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An Information Retrieval Approach to Short Text Conversation. *CoRR* abs/1408.6988 (2014). arXiv:1408.6988

Dacher Keltner and Jonathan Haidt. 2001. Social functions of emotions. (2001).

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980

Post	Model		Response
你来不来撒 Are you coming or not?	Gold		你晓得我的技术撒，晕死。 You know my level, damn
	Seq2Seq		我现在已经在被窝里了 I'm in bed now.
	Specified/ECM	Like	我喜欢去海边。/ 来呀，欢迎欢迎，热烈欢迎 I love to go to the beach. / Come on, welcome. Warm welcome
		Sad	我木有心情啊，我也想去旅游/ 我怕我去不了啊 I'm not in the mood. I want to travel, too. / I am afraid I can't go
		Disgust	怎么指望你们来了? / 不去了，你不去 How do I expect you to come? / You don't go. I won't go either
		Angry	去了就去嘛！！ / 我怕你来, 不来我就来 If I go, I'll go!! / I'm afraid you're coming. I'll come if you don't come
		Happy	恩，我现在在屋头奋战/ 哈哈，我在外面，你来不来嘛 Well, I'm fighting in the house now. / Haha, I am outside. Will you come
	Unspecified	MaxDis	来不了，我现在心情很差！ No, I'm in a bad mood!
		MinDis	我也想啊，可是没人陪我去啊 I'd like to, but no one will go with me.
		MaxEmo	我现在胃痛的很 I have a bad stomachache now.
这个早餐吃的我好腻 I'm sick of this breakfast.	Gold		姐真心觉得好恰的很啊！… I really feel good about it.
	Seq2Seq		我也是，我也是第一次吃，哈哈！ Me too. It's my first time, haha!
	Specified/ECM	Like	我喜欢吃，超级喜欢 / 我喜欢吃的，哈哈 I like it, and very like it / I like to eat, haha
		Sad	我也是啊，可是我的胃也不舒服 / 我也很怀念啊 Me too, but my stomach isn't feeling well. / I miss it, too
		Disgust	那你吃了没? 口味这么重? / 你怎么不说我是吃货呢 Have you eaten yet? The taste is so heavy? / Why don't you say I'm a foodie
		Angry	你以为你是二姐啊？！/ 你是故意的！！ Do you think you are the sister?! / You did it on purpose!!
		Happy	嗯，味道不错，不过吃起来很爽/ 哈哈，我也是，我也超爱吃 Well, it tastes good, but it's very nice to eat. / Haha, me too. I like to eat
	Unspecified	MaxDis	太甜了，我还没吃呢 It's too sweet so that I haven't eaten yet
		MinDis	你吃得太饱了 You've eaten too much
		MaxEmo	我喜欢吃酸辣粉 I like to eat hot and sour rice noodles

Fig. 4: Generated examples using different models. Colored words represent emotion words.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. 110–119.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A Simple, Fast Diverse Decoding Algorithm for Neural Generation. *CoRR* abs/1611.08562 (2016). arXiv:1611.08562

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 1192–1202.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 2157–2169.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-attentive Sentence Embedding. *CoRR* abs/1703.03130 (2017). arXiv:1703.03130

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. 1045–1048.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. 3111–3119.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. 3349–3358.

Yehong Peng, Yizhen Fang, Zhiwen Xie, and Guangyou Zhou. 2019. Topic-enhanced emotional conversation generation with attention mechanism. *Knowl.-Based Syst.* 163 (2019), 429–437. https://doi.org/10.1016/j.knosys.2018.09.006

Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka, and Kenji Araki. 2009. Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*. 1469–1474.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*. 583–593.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical Neural Network Generative Models for Movie Dialogues. *CoRR* abs/1507.04808 (2015). arXiv:1507.04808

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. 1577–1586.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating Long and Diverse Responses with Neural Conversation Models. *CoRR* abs/1701.03185 (2017). arXiv:1701.03185

Marcin Skowron. 2009. Affect Listeners: Acquisition of Affective States by Means of Conversational Systems. In *Development of Multimodal Interfaces: Active Listening and Synchrony, Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*. 169–181. https://doi.org/10.1007/978-3-642-12397-9_14

Marcin Skowron, Stefan Rank, Mathias Theunis, and Julian Sienkiewicz. 2011. The Good, the Bad and the Neutral: Affective Profile in Dialog System-User Communication. In *Affective Computing and Intelligent Interaction - 4th International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part I*. 337–346. https://doi.org/10.1007/978-3-642-24600-5_37

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. 196–205.

Xiao Sun, Xiaoqi Peng, and Shuai Ding. 2017. Emotional Human-Machine Conversation Generation Based on Long Short-Term Memory. *Cognitive Computation* (2017), 1–9.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 3104–3112.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. *CoRR* abs/1701.03079 (2017). arXiv:1701.03079

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to Make Context More Useful? An Empirical Study on Context-Aware Neural Conversational Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*. 231–236. https://doi.org/10.18653/v1/P17-2036

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *CoRR* abs/1610.02424 (2016). arXiv:1610.02424

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A Dataset for Research on Short-Text Conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest*

Group of the ACL. 935–945.

Jason D. Williams and Steve J. Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language* 21, 2 (2007), 393–422. https://doi.org/10.1016/j.csl.2006.06.008

Zhiwen Xie, Zhao Zeng, Guangyou Zhou, and Weijun Wang. 2017. Topic enhanced deep structured semantic models for knowledge base question answering. *SCIENCE CHINA Information Sciences* 60, 11 (2017), 110103:1–110103:15. https://doi.org/10.1007/s11432-017-9136-x

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 3351–3357.

Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 685–694. https://doi.org/10.1145/3077136.3080843

Rui Zhang, Zhenyu Wang, and Dongcheng Mai. 2017. Building Emotional Conversation Systems Using Multi-task Seq2Seq Learning. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*. 612–621. https://doi.org/10.1007/978-3-319-73618-1_51

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. 250–259. http://aclweb.org/anthology/P/P15/P15-1025.pdf

Guangyou Zhou and Jimmy Xiangji Huang. 2017. Modeling and Learning Distributed Word Representation with Metadata for Question Retrieval. *IEEE Trans. Knowl. Data Eng.* 29, 6 (2017), 1226–1239. https://doi.org/10.1109/TKDE.2017.2665625

Guangyou Zhou, Zhiwen Xie, Jimmy Xiangji Huang, and Tingting He. 2016b. Bi-Transferring Deep Neural Networks for Domain Adaptation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. http://aclweb.org/anthology/P/P16/P16-1031.pdf

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. *CoRR* abs/1704.01074 (2017). arXiv:1704.01074

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016a. Multi-view Response Selection for Human-Computer Conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 372–381.

Yimeng Zhuang, Xianliang Wang, Han Zhang, Jinghui Xie, and Xuan Zhu. 2017. An Ensemble Approach to Conversation Generation. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*. 51–62. https://doi.org/10.1007/978-3-319-73618-1_5